

Inference from Private Microdata

Debolina Ghatak

TCG Crest, Kolkata

March 09, 2022

Problem Description

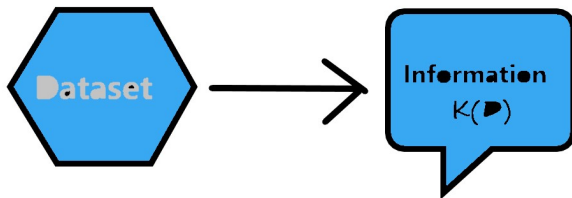
- ▶ Microdata are very important part of industrial and research work.
- ▶ A typical data-set consists of n individuals, m attributes corresponding to each individual.
- ▶ Privacy Issues may arise due to disclosure of raw statistical data-sets.

Example: Insurance data-set

A male senior citizen residing at southeast region who has 3 children. Only one row having such information. Charges and bmi identified.

age	sex	bmi	children	smoker	region	charges
54	female	31.9	1	no	southeast	10928.849
37	male	37.07	1	yes	southeast	39871.7043
63	male	31.445	0	no	northeast	13974.45555
21	male	31.255	0	no	northwest	1909.52745
54	female	28.88	2	no	northeast	12096.6512
60	female	18.335	0	no	northeast	13204.28565
32	female	29.59	1	no	southeast	4562.8421
47	female	32	1	no	southwest	8551.347
21	male	26.03	0	no	northeast	2102.2647
28	male	31.68	0	yes	southeast	34672.1472
63	male	33.66	3	no	southeast	15161.5344
18	male	21.78	2	no	southeast	11884.04858
32	male	27.835	1	no	northwest	4454.40265
38	male	19.95	1	no	northwest	5855.9025
32	male	31.5	1	no	southwest	4076.497
62	female	30.495	2	no	northwest	15019.76005
.		

Question of security



Secure or not??

Privacy guarantees to Identity Disclosure

The identity disclosure problem deals with ensuring that the **identity of individuals cannot be guessed with the information released**. One of the most well-known privacy measures was:

- ***k*-anonymity** proposed by Sweeney (2002)[Swe02].

However, there were many challenges dealing with it. Two famous attacks were proposed to this guarantee including remedies like adding the guarantees of

- ***l*-diversity** : proposed by Machanavajjhala et. al.(2006)[MGKV06]
- ***t*-closeness** : proposed by Li et. al. (2007) [LLV07]

Differential Privacy

Due to too many restrictions to ensure k -anonymity, l -diversity and t -closeness to a released data-set, it was hardly useful in any practical situation. There was a huge struggle for scientists and practitioners to give one precise privacy guarantee to a released statistical information until the advent of the idea of differential privacy.

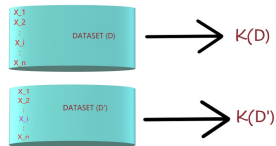
Differential Privacy :

The privacy policy as proposed by Dwork et. al. (2006) [DMNS06] guarantees that for every subrange of the released information the likelihood of it coming from a data-set containing information of any row x and that of not containing the information of x is nearly the same .

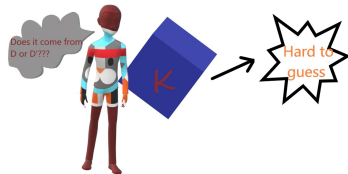
Thus an intruder cannot identify from released information whether his target individual is in the data-set or not.

Theory of Differential Privacy

A mechanism \mathcal{K} to be applied on a dataset D is said to be differentially private if the output for any two neighbouring data-sets has approximately same probability to take certain value(s).



Thus, an attacker looking at the released information from the data-set would be confused if his target individual is present/absent in the data-set.



Mathematical Formulation

Mathematically it compares the probability of any released information $\mathcal{K}(\cdot)$ coming from a data-set D with that of one coming from a neighbouring data-set of it and restricts it's value to be close to 1. To ensure,

$$\frac{P[\mathcal{K}(D) \in B]}{P[\mathcal{K}(D') \in B]} \leq e^\epsilon$$

where $D \sim D'$ are two data-sets differing in one row, $\epsilon > 0$ is small and $B \subset \text{Range}(\mathcal{K})$.

Justification of Noise Addition to achieve Differential privacy

Dwork et. al. (2006) [DMNS06] develops a theorem that justifies that Laplace Noise Addition to bounded functions can help in achieving Differential Privacy guarantee to information release.

Statement

The L_1 -sensitivity of a function $f : \mathcal{D} \rightarrow R^d$ can be defined as

$$\nabla f = \max_{D \sim D'} \|f(D) - f(D')\|$$

Then,

$$f(D) + Y \text{ where, } Y = (Y_1, Y_2, \dots, Y_d) \text{ and } Y_i \sim \text{Lap}(\nabla f / \epsilon)$$

satisfies ϵ -DP guarantee.

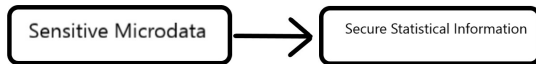
NOTE: Other mechanisms have also been proposed later on to satisfy this guarantee for example K-norm mechanism, Exponential mechanism.

Some Variants of Differential Privacy

Once Differential Privacy guarantee was proposed, researchers came up with several similar privacy measures which are also of great use. Some of them are as follows:

- Random Differential Privacy
- Rényi Differential Privacy
- Approximate Differential Privacy
- Local Differential Privacy

Initial Statistical Inference Technique



Examples

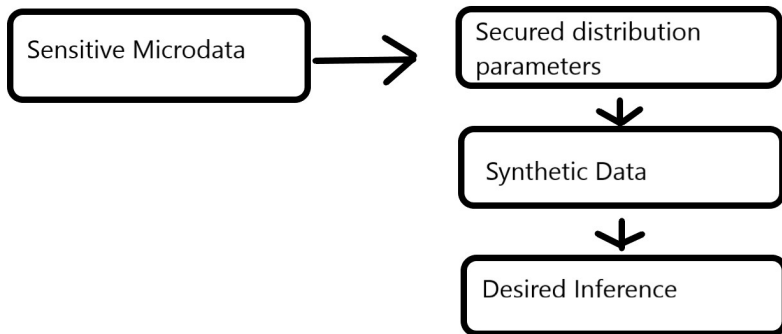
- Differentially private mean
- Differentially private median
- Differentially private histogram counts, etc.

Private Microdata Release

Synthetic Data Generation

A synthetic data-set is a data-set that is not true data of any individual but behaves statistically like true data-set.

Recent Statistical Inference Technique



Some Pioneer Works

Categorical attributes

- For the low dimensional case, one can form **contingency tables**, add Laplace noise to guarantee differential privacy, generate synthetic data from it.
- For high dimensional case, one needs to assume or test the dependence structure among attributes to reduce complexity of the procedure and then perform the same procedure.

Numerical attributes

- For the low dimensional case, one can form **histograms**, add Laplace noise to the bin counts to guarantee differential privacy, generate synthetic data from it using rejection sampling. Another method is to use **Generative Adversarial Networks (GAN)** by putting noise into the stochastic gradient method.
- For the high-dimensional case, one method to deal with it is to assume a **Copula structure** for the data and execute the algorithms.

Privacy vs Utility

- Privacy is very important but so is utility of available data after guaranteeing privacy to it.
- Differential privacy theoretically guarantees privacy and justifies use of noise addition to ensure privacy.
- But coming into a more practical scenario, a question still remains about the choice of parameters that might take care of both privacy and utility.

Choice of privacy and utility parameters

What is ϵ in ϵ -DP ??

Muralidhar et al. (2020) [MDFM20] experimentally shows that the parameter ϵ in ϵ -differential privacy is neither a measure of confidentiality nor utility and hence can be misleading in understanding how much confidentiality is ensured due to a particular choice of it. The significance of the choice of parameter is important to understand but its study has been neglected.

The problem of Data Obfuscation

The problem of data obfuscation aims to

- Minimize the disclosure risk of sensitive data
- Maximize the utility of private data

Moving to the privacy-utility tradeoff formulation

- A few works have been done by Rastogi et al (2007)[RHS07], Ghatak and Roy (2018)[GR18] in this domain.
- Salamatian et al. (2020) [SCF⁺20] views the problem in an information theoretic approach and comes up with the concept of Privacy Funnel.

Privacy Funnel

Break available data into sensitive variables (S) and useful variables (X). Release Y .

Set-Up

$$(S, X) \longrightarrow Y \longrightarrow \hat{S} \text{ (privacy)}$$

$$(S, X) \longrightarrow Y \longrightarrow \hat{X} \text{ (utility)}$$

Assumptions

$$X \sim \mathcal{X} = [m], Y \sim \mathcal{Y} = [n],$$

$$\Delta_m = \{x \in \mathbb{R}^m \mid \sum_{i=1}^m x_i = 1, x_i \geq 0\}$$

and similarly Δ_n denotes column probability domain for X and Y .

T : Channel transformation from X to Y

$$q = Tp$$

Optimization

The privacy funnel chooses the joint entropy between X and Y as a metric of information and finds an optimal transformation map that solves,

$$\text{Minimize } I(S, Y)$$

$$p_{Y|X} : I(X, Y) \geq t$$

for a given utility level t for $0 \leq t \leq H(X)$ (the entropy of X).






Open Problems for Synthetic Data Generation

- Some challenges for statisticians include developing practically useful methods of secured synthetic data generation from multivariate empirical distribution functions.
- For mixed data sets containing both categorical (ordinal and nominal) and continuous attributes, the generation of synthetic data will tend to be harder. This might be a challenging problem in this field of work.

Open problems in Information theoretic approach

- One of the most significant challenges in this field is perhaps to develop meaningful frameworks to achieve the privacy-utility tradeoff for general microdata which might be useful in practical scenario.
- Sticking to the noise addition model, developing ideas to find ideal parameters is also a practical challenge.

References I

-  C. Dwork, F. McSherry, K. Nissim, and A. Smith., *Calibrating noise to sensitivity in private data analysis*, In Proceedings of the Third Conference on Theory of Cryptography, 2006, p. 265–284.
-  D. Ghatak and B. Roy, *Estimation of true quantiles from quantitative data obfuscated with additive noise*, Journal of Official Statistics, 2018.
-  N. Li, T. Li, and S. Venkatasubramanian, *t-closeness: Privacy beyond k-anonymity and l-diversity*, IEEE 23rd International Conference on Data Engineering, Istanbul, 2007, pp. 106–115.
-  Krish Muralidhar, Josep Domingo-Ferrer, and Sergio Martínez, *epsilon-differential privacy for microdata releases does not guarantee confidentiality (let alone utility)*, In book: Privacy in Statistical Databases, UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, Spain, September 23–25, 2020, Proceedings, 2020.
-  A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian, *L-diversity: privacy beyond k-anonymity*, 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 2006.

References II



Vibhor Rastogi, Sungho Hong, and Dan Suciu, *The boundary between privacy and utility in data publishing.*, 09 2007, pp. 531–542.



Salman Salamatian, F. Calmon, N. Fawaz, A. Makhdoumi, and M. Médard, *Privacy-utility tradeoff and privacy funnel.*



L. Sweeney, *k-anonymity: A model for protecting privacy.*, vol. 10:5, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002, pp. 557—570.

Thank You